PATENT ABSTRACTS OF JAPAN

(11)Publication number : **11-352994**
(43)Date of publication of application : **24.12.1999**

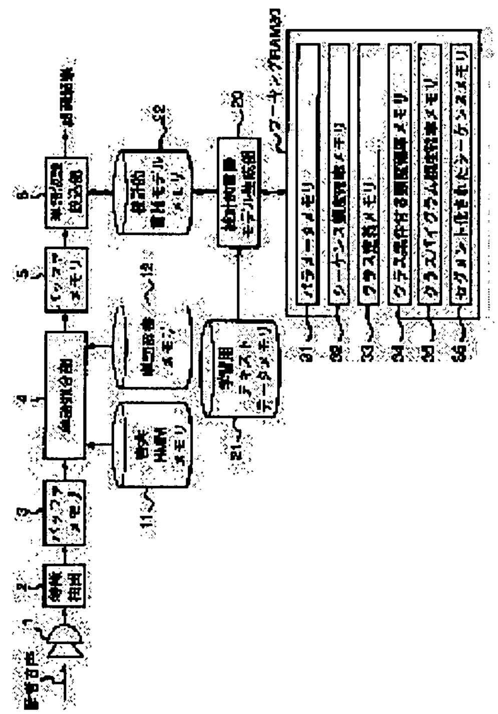
(51)Int.CI. G10L 3/00
G10L 3/00
G10L 3/00
// C12N 15/09

(21)Application number : **10-165030**
(22)Date of filing : **12.06.1998**

(71)Applicant : **ATR ONSEI HONYAKU TSUSHIN KENKYUSHO:KK**
(72)Inventor : **SABIN DERIN**
KOSAKA YOSHINORI
NAKAJIMA HIDEJI

(54) STATISTICAL SEQUENCE MODEL GENERATOR, STATISTICAL LANGUAGE MODEL GENERATOR, AND SPEECH RECOGNITION SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To impart degree of freedom to an analyzed result, and to process a variable length of sequence as the same class, by re-estimating frequency probability of a bigram between sequences based on front likelihood, frequency probability, and rear likelihood.
SOLUTION: Re-estimation is conducted in a statistical language model generating part 20 to provide the maximum likelihood estimate using EM algorithm based on frequency probability for character lines included in sorted classes and conditional character lines and frequency probability for a bigram between the classes. Frequency probability of a bigram between sequences is re-estimated using an expression for expressing the frequency probability of the bigram between the sequences based on front likelihood for the character lines of processing objects put in a front side of a time series, frequency probability of the character lines when the character line just before the character lines is conditioned, and rear likelihood corresponding to the character lines put in a rear side of the time series, relating to the respective character lines of the processing objects. Bi-multigram statistical sequence models are generated to be output.

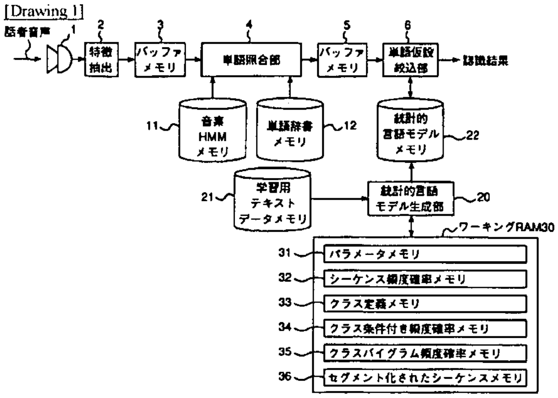


LEGAL STATUS

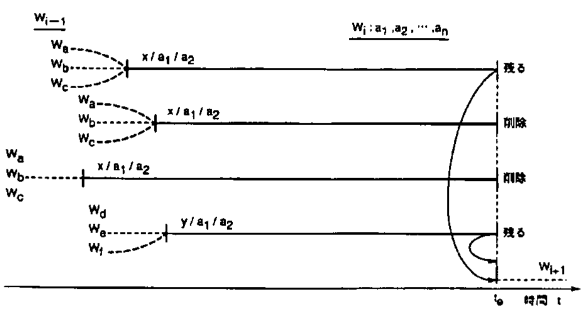
[Date of request for examination]	12.06.1998
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	3004254
[Date of registration]	19.11.1999
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	

DRAWINGS

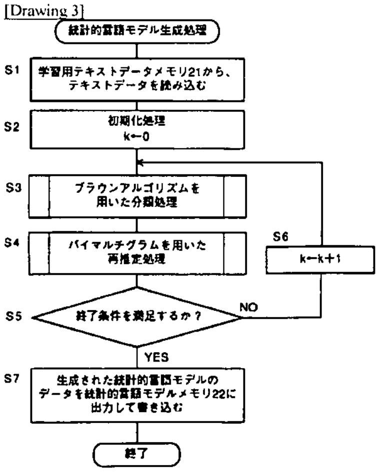
[Drawing 1]



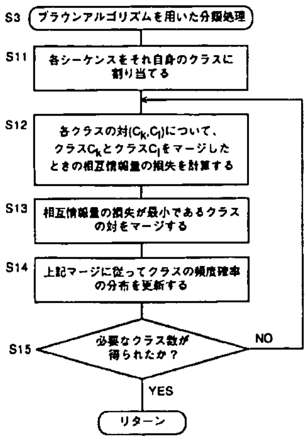
[Drawing 2]



[Drawing 3]



[Drawing 4]



[Translation done.]

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is the block diagram of the continuous speech recognition equipment which is 1 operation gestalt concerning this invention.
[Drawing 2] It is the timing chart which shows processing of the word hypothetical narrowing-down section 6 in the continuous speech recognition equipment of drawing 1 .
[Drawing 3] It is the flow chart which shows the statistical language model generation processing performed by the statistical language model generation section 20 of drawing 1 .
[Drawing 4] It is the flow chart which shows the classification processing using the Brown algorithm which is the sub routine of drawing 3 .
[Description of Notations]
1 -- Microphone
2 -- Feature-extraction section,
3 5 -- Buffer memory
4 -- Word collating section,
6 -- Word hypothetical narrowing-down section,
11 -- Phoneme HMM memory,
12 -- Word dictionary memory,
20 -- Statistical language model generation section,
21 -- Text data memory for study,
22 -- Statistical language model memory,
30 -- Working RAM
31 -- Parameter memory,
32 -- Sequence frequency probability memory,
33 -- Class definition memory,
34 -- Frequency probability memory with class condition,
35 -- Class motorcycle gram frequency probability memory,
36 -- Segmented sequence memory.

[Translation done.]

PAT-NO: JP411352994A

DOCUMENT-IDENTIFIER: JP 11352994 A

TITLE: STATISTICAL SEQUENCE MODEL GENERATOR, STATISTICAL LANGUAGE MODEL GENERATOR, AND SPEECH RECOGNITION SYSTEM

PUBN-DATE: December 24, 1999

INVENTOR-INFORMATION:

NAME	COUNTRY
SABIN, DERIN	N/A
KOSAKA, YOSHINORI	N/A
NAKAJIMA, HIDEJI	N/A

ASSIGNEE-INFORMATION:

NAME	COUNTRY
ATR ONSEI HONYAKU TSUSHIN KENKYUSHO:KK	N/A

APPL-NO: JP10165030

APPL-DATE: June 12, 1998

INT-CL (IPC): G10L003/00, G10L003/00 , G10L003/00 , C12N015/09

ABSTRACT:

PROBLEM TO BE SOLVED: To impart degree of freedom to an analyzed result, and to process a variable length of sequence as the same class, by re-estimating frequency probability of a bigram between sequences based on front likelihood, frequency probability, and rear likelihood.

SOLUTION: Re-estimation is conducted in a statistical language model generating part 20 to provide the maximum likelihood estimate using EM algorithm based on frequency probability for character lines included in sorted classes and conditional character lines and frequency probability for a bigram between the classes. Frequency probability of a bigram between sequences is re-estimated using an expression for expressing the frequency probability of the bigram between the sequences based on front likelihood for the character lines of processing objects put in a front side of a time series, frequency probability of the character lines when the character line just before

the character lines is conditioned, and rear likelihood corresponding to the character lines put in a rear side of the time series, relating to the respective character lines of the processing objects. Bi-multigram statistical sequence models are generated to be output.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[The technical field to which invention belongs] this invention relates to the voice recognition unit which carries out speech recognition of the sound signal of the phonation voice sentence inputted using the statistical sequence model generation equipment which generates a statistical sequence model based on the sequence data for study, the statistical language model generation equipment which generates a statistical language model based on the text data for study, and the above-mentioned statistical language model.
[0002]
[Description of the Prior Art] In recent years, in order to raise the performance in continuous speech recognition equipment, the method of using a language model is studied. This aims at the improvement in a recognition rate, and the effect of curtailment of machine time by predicting the following word and cutting down the search space using the language model which is a sequence model. Here, specifically, a sequence is a word in the sequence of a character and is a phrase (or phrase) in the sequence of a word. there is N-gram (N-gram; -- here, N is the two or more natural numbers) as a language model used briskly recently This learns large-scale text data and gives statistically the transition probability from the last word to the N-1 word as follows. Two or more generation probability P of L word train w1 L=w1, and w2, --, wL (w1L) is expressed with the following formula.
[0003]
[Equation 1]

$$P(w_1^L) = \prod_{t=1}^{L} P(w_t \mid w_{t+1-N}^{t-1})$$

[0004] Here, wt expresses the t-th one word among word train w1L, and wij expresses the i-th to the j-th word train. In one above, Probability P (wt|wt+1-Nt-1) is the probability that the word "wt" will be uttered, after word train wt+1-Nt-1 which consists of N words is uttered, and like the following, Probability P (A|B) means the probability that the word "A" will be uttered, after a word or the word train B is uttered. Moreover, "pi" in several 1 means the product of the probability P from t= 1 to L (wt|wt+1-Nt-1).
[0005] By the way, the technique of raising the performance of a continuous speech recognition using N-gram of the above-mentioned statistical language model is proposed briskly, and the method of using the dependency of the word covering the word train of variable length is used with some of the models in recent years. These models are used in order to ease assumption of the dependency of the fixed length seen by the conventional N-gram model in common, and they cover various larger assumption.
[0006] In order to draw a phrase purely by the statistical method (namely, method of not using a grammatical rule which is in a statistical context free language (Stochastic Context Free Grammars)), various criteria needed to be used, for example, the following criteria have been proposed.
(a) In the conventional technical reference 1 "K.Ries et al., "Class phrase models for languagemodeling", Proceedings of ICSLP 96, and 1996" The indicated reeve one out (leave-one-out) likelihood, "And the (b) conventional technical reference 2 H.Masataki et al. and Variable-order n-gram generation by word-class splitting and consecutive word grouping.Proceedings ofICASSP 96, entropy indicated in 1996."
[0007]
[Problem(s) to be Solved by the Invention] In these methods, although the method of optimization using EM (Expectation Maximum;, i.e., maximization of expected value) algorithm can be used by using the criteria of a likelihood in a statistical framework, there is an inclination used as fault study. Moreover, in optimization processing, although heuristic programming is used in the conventional technical reference 3 "S.Matsunaga et al., "Variable-length language modeling integrating global constraints", Proceedings of EUROSPEECH 97, and 1997", convergence and optimization of a statistical language model are not guaranteed theoretically, for example.
[0008] It is as follows when the trouble when using further the criteria of the likelihood proposed in the conventional technical reference 1 here is described.
Since the frequency probability of the sequence of a <trouble 1> word is acquired by the avaricious algorithm (greedy algorithm), monotonous convergence which tends toward the optimal state is not guaranteed.
<Trouble 2> This method is deterministic. That is, if a sequence [bcd] is in the TOC (inventory) of a sequence, this will not be divided into subsequences, such as [bc] + [d] and [b]+ [cd] and [b]+[c]+ [d], even if "bcd" occurs in an input character string. In other words, there is no flexibility in the analysis to a sequence.
It is based on the class classification of the word which the definition of the class of a <trouble 3> sequence precedes. That is, it is

used, in order that a word may be classified, next each sequence of the label of the class of a word may define the class of a sequence first. Therefore, the sequence from which length is different cannot be put into the same class. For example, "thank you for" and "thank you very much for" do not go into the same class.

[0009] In order to solve this, this invention person "The conventional technical reference 4 S.Deligne et al., "Introducing statistical dependencies and structural constraints in variable-length sequence models", In Grammatical Inference: In Learning Syntaxfrom Sentences, Lecture Notes in Artificial Intelligence 1147, pp.156-167, Springer, and 1996" Although only possibility that those parameters will be calculable using (16) formulas of the conventional technical reference 4 concerned about the statistical language model using the multi-gram which is the sequence of variable length is shown The (16) formulas concerned did not turn into a form actually calculable using a digital computer, but had the trouble of being unutilizable. Here, a multi-gram is the sequence of the variable length which does not specify a dependency with other sequences.

[0010] The purpose of this invention can solve the above trouble, and can guarantee monotonous convergence which tends toward the optimal state as compared with the conventional example, flexibility is in an analysis result, and it is in offering the statistical sequence model generation equipment, the statistical language model generation equipment, and the voice recognition unit which can deal with the sequence of variable length in the same class, can carry out high-speed processing practical using a digital computer, and can generate a statistic model.

[0011]

[Means for Solving the Problem] The statistical sequence model generation equipment concerning this invention It is based on input data including the sequence which is the unit string which consists of one piece or two or more units. It is statistical sequence model generation equipment which generates the statistical sequence model of the Bayh-multi-gram which is BAIGURAMU between the natural number N1 piece unit string of variable length, and the natural number N2 piece unit string of variable length. Based on the above-mentioned input data, under the constraint of the maximum of N1 and N2 which were decided beforehand The initialization means which carries out counting of the frequency probability of the above-mentioned motorcycle gram of the combination of all unit strings, It is based on the frequency probability of the above-mentioned motorcycle gram in which counting was carried out by the above-mentioned initialization means. By classifying into a number of two or more classes which merged so that loss of the mutual information when merging the pair of each class might serve as the minimum, updated the frequency probability of each class, and were decided beforehand The frequency probability of the unit string contained in the classified class, and the conditional unit string of the classified class, A classification means to calculate and output the frequency probability of BAIGURAMU between the classified classes, The unit string contained in the classified class which is outputted from the above-mentioned classification processing means, Based on the frequency probability of the conditional unit string of the classified class, and the frequency probability of BAIGURAMU between the classified classes, it is re-presumed using EM algorithm that a maximum likelihood estimate is obtained. here Each unit string of a processing object is received using a forward back WORD algorithm. The front likelihood to the unit string concerned of the processing object serially obtained for the front, The frequency probability of the unit string concerned at the time of being contingent [on the unit string in front of the unit string concerned], By re-presuming the frequency probability of BAIGURAMU between the sequences concerned using the formula showing the frequency probability of BAIGURAMU between sequences based on the back likelihood to the unit string concerned to obtain for back serially It is characterized by having a re-presumption means to generate and output the statistical sequence model of the above-mentioned Bayh-multi-gram which it is as a result of re-presumption, and the control means controlled to perform processing of the above-mentioned classification means, and processing of the above-mentioned re-presumption means repeatedly until it fulfills predetermined end conditions.

[0012] Moreover, in the above-mentioned statistical sequence model generation equipment, the above-mentioned initialization means is characterized by removing the data of the combination of BAIGURAMU below predetermined frequency probability further among the frequency probability of BAIGURAMU by which counting was carried out [above-mentioned].

[0013] Furthermore, in the above-mentioned statistical sequence model generation equipment, the above-mentioned classification means is characterized by classifying into two or more above-mentioned classes according to the above-mentioned initialization means using the Brown algorithm based on the frequency probability of the above-mentioned motorcycle gram by which counting was carried out.

[0014] In the above-mentioned statistical sequence model generation equipment moreover, the above-mentioned formula In the above-mentioned input data, the frequency probability of BAIGURAMU between the sequences of a unit string in case the 2nd unit string which is the unit string concerned follows the 1st unit string It is a formula for calculating to each unit string of the processing object in the above-mentioned input data. the frequency probability of BAIGURAMU between the above-mentioned sequences It is obtained by carrying out the division of the sum of the likelihood in all segmentation containing the 1st and the 2nd unit string by the sum of the likelihood in all segmentation containing the 1st unit string. Moreover, the denominator which shows the number of times of an average to which each unit string generates the above-mentioned formula in the above-mentioned input data here, It has the molecule which shows the number of times of an average to each unit string in case the 2nd unit string follows the 1st unit string in the above-mentioned input data. the above-mentioned molecule It is the sum of the above-mentioned front likelihood to each unit string of a processing object, the frequency probability of the unit string concerned at the time of being contingent [on the unit string in front of the unit string concerned], and the product of the above-mentioned back likelihood. the above-mentioned denominator It is the sum of the above-mentioned front likelihood to each unit string of a processing object, the frequency probability of all the unit strings at the time of being contingent [on the unit string in front of the unit string concerned], and the product of the above-mentioned back likelihood.

[0015] Furthermore, in the above-mentioned statistical sequence model generation equipment, the above-mentioned end conditions are characterized by being a time of reaching the number of times the number of occurrence of processing of the above-mentioned classification means and processing of the above-mentioned re-presumption means was beforehand decided to be.

[0016] Moreover, in the above-mentioned statistical sequence model generation equipment, the above-mentioned unit of the statistical

language model generation equipment concerning this invention is the character of natural language, the above-mentioned sequence is a word, the above-mentioned classification means classifies a character string into the train of two or more words, and the above-mentioned statistical sequence model is characterized by being a statistical language model.
[0017] Furthermore, in the above-mentioned statistical sequence model generation equipment, the above-mentioned unit of the statistical language model generation equipment concerning this invention is the word of natural language, the above-mentioned sequence is a phrase, the above-mentioned classification means classifies a word train into the train of two or more phrases, and the above-mentioned statistical sequence model is characterized by being a statistical language model.
[0018] Furthermore, based on the sound signal of the phonation voice sentence into which the voice recognition unit concerning this invention is inputted, the above-mentioned speech recognition means is characterized by carrying out speech recognition with reference to the statistical language model generated by the above-mentioned statistical language model generation equipment in the voice recognition unit equipped with the speech recognition means which carries out speech recognition using a predetermined statistical language model.
[0019]
[Embodiments of the Invention] Hereafter, the operation gestalt which starts this invention with reference to a drawing is explained. In the following operation gestalten, although an example which classifies into a phrase (phrase) the word train which an example which classifies into a word train the character string which a unit is a character and is the sequence of a character, and a unit are words, and is the sequence of a word is explained, you may constitute so that not only this but the DNA train which it is this invention, and a unit is DNA, and is the sequence of DNA may be classified into a predetermined DNA array. Moreover, a unit is a base, and you may constitute it so that the base train which is the sequence of a base may be classified into a predetermined codon.
[0020] Drawing 1 is the block diagram of the continuous speech recognition equipment which is 1 operation form concerning this invention. The continuous speech recognition equipment of this operation form is based on the text data which is the character string memorized by the text data memory 21 for study. It has the statistical language model generation section 20 which generates the language model of the Bayh-multi-gram of variable length using working RAM 30. here processing of the statistical language model generation section 20 If it roughly divides as shown in drawing 3 , it is characterized by including the classification processing (Step S3) using the Brown algorithm, and re-presumption processing (Step S4) in which the Bayh-multi-gram was used.
[0021] Namely, the statistical language model generation equipment of this operation gestalt It is based on input data including the sequence of a character string which consists of one piece or two or more characters. It is statistical language model generation equipment which generates the statistical language model of the Bayh-multi-gram which is BAIGURAMU between the natural number N1 piece character string of variable length, and the natural number N2 piece character string of variable length. here As shown in drawing 3 , based on the (a) above-mentioned input data, under the constraint of the maximum of N1 and N2 which were decided beforehand The initialization processing which carries out counting of the frequency probability of the above-mentioned motorcycle gram of the combination of all character strings (Step S2), (b) It is based on the frequency probability of the above-mentioned motorcycle gram in which counting was carried out by the above-mentioned initialization processing. By classifying into a number of two or more classes which merged so that loss of the mutual information when merging the pair of each class might serve as the minimum, updated the frequency probability of each class, and were decided beforehand The frequency probability of the character string contained in the classified class, and the conditional character string of the classified class, The classification processing which calculates and outputs the frequency probability of BAIGURAMU between the classified classes (Step S3), (c) The character string contained in the classified class which was obtained by the above-mentioned classification processing, Based on the frequency probability of the conditional character string of the classified class, and the frequency probability of BAIGURAMU between the classified classes, it is re-presumed using EM algorithm that a maximum likelihood estimate is obtained. here Each character string of a processing object is received using a forward back WORD algorithm. The front likelihood to the character string concerned of the processing object serially obtained for the front, The frequency probability of the character string concerned at the time of being contingent [on the character string in front of the character string concerned], By re-presuming the frequency probability of BAIGURAMU between the sequences concerned using the formula (several-24 22-number) showing the frequency probability of BAIGURAMU between sequences based on the back likelihood to the character string concerned to obtain for back serially The re-presumption processing which generates and outputs the statistical sequence model of the above-mentioned Bayh-multi-gram which it is as a result of re-presumption (step S4), (d) It is characterized by including the processing (Step S5) controlled to perform the above-mentioned classification processing and the above-mentioned re-presumption processing repeatedly until it fulfills predetermined end conditions.
[0022] This operation gestalt focuss on the method based on a phrase of countering the technique based on N-gram of a word. Here, two or more sentences are constituted by the phrase and frequency probability is assigned to a phrase instead of a word. It is not concerned with whether a model is based on N-gram, or it is based on a phrase, but they correspond to either a deterministic model or a statistic model. By the framework based on a phrase, un-deciding nature is introduced into a phrase through the ambiguity of the analysis result of the sentence. That is, it means that this does not have no probability that the analysis result of a character string will be set to [ab] and [c] in spite of actually setting and registering phrase "abc" as a phrase. this -- contrastive -- the deterministic technique -- a, b, and c -- all simultaneous appearances are systematically interpreted as the appearance of a phrase [abc]
[0023] Moreover, with this operation gestalt, processing of a statistical language model is performed using a Bayh-multi-gram, the language model of the Bayh-multi-gram concerned is a statistic model based on a phrase, and the parameter is presumed in accordance with likelihood criteria.
[0024] First, theoretical formulization of a multi-gram is explained. The sentence which consists of T words by the framework of a multi-gram [several 2] W=w (1) w (2) -- w (T) is assumed to be that in which each phrase which consists of a word of the n maximum length, respectively carried out the chain (sequence). Here, S shows the segmentation to the phrase of Ts individual, and s (t) can express the result of the segmentation by S of W with the following formula, when it considers as the phrase of the time index (the

serial number from the first word is shown.) (t) in segmentation S.
[Equation 3] (W, S) =s (1) -- s(Ts) [0025] Here, the dictionary which consists of two or more segmented phrases is formed combining a word until it results in n from 1 and 2 -- from a vocabulary, and expresses like the following formula.
[Equation 4] Ds= {sj} j and the likelihood of a sentence are calculated like the following formula as the sum of the likelihood to each segmentation.
[0026]
[Equation 5]

$$L(W) = \sum_{S \in \{S\}} L(W, S)$$

[0027] Sentence W is analyzed by the determination inclination-technique of a model according to the segmentation most appropriate for **, and the following approximation is acquired by it.
[0028]
[Equation 6]

$$L^{*}(W) = \max_{S \in \{S\}} L(W, S)$$

[0029] Here, correlation of n-gram between phrases is assumed and the value of the likelihood as a result of specific segmentation S is calculated like the following formula.
[0030]
[Equation 7]

$$L(W, S) = \prod_{t} p(s_{(t)} \mid s_{(t-n+1)} \cdots s_{(t-1)})$$

[0031] Here, hereafter, Sign n expresses the dependence between two or more phrases, and uses it as n of the notation of conventional n-gram. Moreover, Sign nmax expresses the maximum length of a phrase. Therefore, the example of calculation of a likelihood is shown in the following formula here. This example shows the likelihood of "abcd" of a Bayh-multi-gram model (nmax=3, n= 2). Sign # expresses an empty sequence.
[0032]
[Equation 8] Likelihood =p p ([a] |#) p ([b] | [a]) p ([c] | [b]) ([d] | [c]) +p p ([a] |#) p ([b] | [a]) ([cd] | [b]) +p p ([a] |#) p ([bc] | [a]) ([d] | [bc]) +p([a] |#) p([bcd] | [a]) +p([ab] |#) p([c] | [ab]) p([d] | [c]) +p([ab] |#) p([cd] | [ab]) +p([abc] |#) p ([d] | [abc]) [0033] The likelihood concerned expresses the sum of the frequency probability about all the combination when segmenting sequence "abcd" so that clearly from eight above.
[0034] Subsequently, presumption of the parameter of a language model is explained. The n-gram model of a multi-gram is completely defined by the set of Parameter theta, the parameter theta of the following formula uses Dictionary Ds, and it is [Equation 9].
theta= {p(sin|si1 --sin-1) |si1 --sin**Ds}
conditional [of n-gram related to all the combination of n phrases] -- it is constituted by probability It is obtained as the likelihood value of the maximum which can be assumed acquired from imperfect data, i.e., a maximum likelihood estimate, (Maximum Likelihood Estimation), and the estimate of the set of Parameter theta is segmentation S to which strange data make the foundation here. Therefore, the repetitive maximum likelihood estimate of Parameter theta is calculable with well-known EM algorithm (Expectation Maximization Algorithm). Here, let Q (k, k+1) be the auxiliary function of the following formula calculated using the number-of-occurrence parameter k and the likelihood of k+1.
[0035]
[Equation 10]

$$Q(k, k+1) = \sum_{S \in \{S\}} L^{(k)}(S \mid W) \log \{L^{(k+1)}(W, S)\}$$

[0036] It is [Equation 11] as well-known EM algorithm is shown. Q(k, k+1) >=Q (k, k)
It will be [Equation 12], if come out and it is. L (k+1) (W) >=L (k) (W)
It comes out. Therefore, the re-presumption formula of the following formula in a number-of-occurrence parameter (k+1) [several 13] p (k+1) (sin|si1 --sin-1)
The constraint of ** and the following formula [several 14]

$$\sum_{s_{in} \in D_s} p(s_{in} \mid s_{i1} \cdots s_{in-1}) = 1$$

It can lead directly like the following formula by maximizing the auxiliary function Q (k, k+1) about the model parameter theta (k+1) also as **. In addition, in this specification, since the notation with the bottom with the bottom and the notation with the bottom with a top are not made, the notation with the bottom of a lower layer is omitted.
[0037]
[Equation 15]

$$p^{(k+1)}(s_{in} \mid s_{i1} \cdots s_{in-1})$$
$$= p_a / p_b$$

ここで、

$$p_a = \{ \sum_{S \in \{S\}} c(s_{i1} \cdots s_{in-1} s_{in}, S) \times L^{(k)}(S \mid W) \}$$
$$p_b = \{ \sum_{S \in \{S\}} c(s_{i1} \cdots s_{in-1}, S) \times L^{(k)}(S \mid W) \}$$

[0038] Here, c (si1 -- sin, S) is two or more phrases si1 in segmentation S. -- The number of appearances of the combination of sin is shown. The re-presumption formula of several 15 is performed using a forward back WORD algorithm (forward backward algorithm) (henceforth the FB method) so that the detailed after-mentioned may be carried out about a Bayh-multi-gram (n= 2). By the determination-oriented method, a re-presumption formula is simplified like the following formula.
[0039]
[Equation 16]

$$p^{(k+1)}(s_{in} \cdots s_{in-1})$$
$$= \{c(s_{i1} \cdots s_{in-1} s_{in}, S^{*(k)})\} / \{c(s_{i1} \cdots s_{in-1}, S^{*(k)})\}$$

[0040] Here, S* (k) is the analysis result of the sentence which maximizes L (k) and (S|W), and is drawn by the Viterbi (Viterbi) algorithm. A language model is improved in each meaning which increases likelihood L (k) and (W) repeatedly, and, finally it converges to the critical point (probably **, partial maximum). The set of the model parameter theta is initialized using the relative frequency of the combination of all the phrases observed in the corpus for study, i.e., the text data for study.
[0041] Subsequently, clustering (classification processing) of a variable length phrase is explained. According to the conventional technical reference 1, although the model based on a class-phrase attracts attention in recent years, it usually assumes the conventional word clustering. Typically, each word can assign first the label Ck of a class with which a word belongs, and the variable length phrase [Ck1, Ck2 --Ckn] of a word-class label is drawn. The label of the class to which the phrase shown as "<[Ck1, Ck2 --Ckn] >" belongs by each variable length phrase is defined. However, the same phrase-class label can be assigned only to the phrase of the same length by this technique. For example, a phrase called "thank you for" and "thank you very much for" cannot be assigned to the same class label. With this operation gestalt, the method of clustering a direct phrase instead of a word as solution over such a limitation is proposed. In order to attain this purpose, correlation (nmax=2) of BAIGURAMU between two phrases is assumed, and change is added to the learning technique of the Bayh-multi-gram model mentioned above, and it is made to consist of the following two stages repeatedly [each].
[0042] (I) Step SS1:class assignment (it corresponds to Step S3 of drawing 3 .)
[Equation 17] {p (k) (sj|si)} -> {p (k), (Ck(sj) |Ck(sj)), p (k) (sj|Ck (sj))}
(II) Re-presumption of a step SS2:multi-gram (it corresponds to Step S4 of drawing 3 .)
[Equation 18] {p (k), (Ck(sj) |Ck(si)), p(k) (sj|Ck (sj))} -> {p (k+1) (sj|si)}
[0043] At the above-mentioned step SS 1, frequency probability of a phrase motorcycle gram is considered as an input, and the frequency probability of a class motorcycle gram is outputted. According to conventional technical reference 5"P.F.Brown et al., "Class-based n-gram models of natural language", Computational Linguistics, Vol.18, No.4, pp.467-479, and 1992", class assignment is performed by maximizing the correlation information between adjacent phrases. Here, the candidate of clustering is taken as the phrase instead of a word. As mentioned above, {p (0) (sj|si)} is initialized using the relative frequency of a simultaneous appearance of the phrase in the text data for study. At the above-mentioned step SS 2, the frequency probability of a phrase is re-presumed using the re-presumption formula (several 15) of a multi-gram, or its approximation (several 16). Here, it is calculated by the formula of the following [difference / only / likelihood / of an analysis result].
[0044]
[Equation 19]

$$L(W, S)$$
$$= \prod_{t} p(C_{k(s(t))} \mid C_{k(s(t-1))})\, p(s_{(t)} \mid C_{k(s(t))})$$

[0045] This is equal to re-presuming the frequency probability p (k+1) (sj|si) based on frequency probability p(k) (Ck(sj) |Ck(si)) xp (k) and (sj|Ck(sj)) like the processing to frequency probability p (k) and (sj|si), as mentioned above.
[0046] If it summarizes, it will be guaranteed that the class assignment based on the criteria of mutual information is optimized by the above-mentioned step SS 1 about the present phrase distribution, and it will be guaranteed that the likelihood calculated by the above-mentioned step SS 2 according to 19 above using the frequency probability of the present class is optimized by the frequency probability of a phrase. Study data follow and are repetitively constituted in the both sides of integrated union level (paradigmatic) (syntagmatic) (it is the term of linguistics, respectively.) by the method integration-ized completely. That is, the union-relation between the phrases expressed by class assignment affects re-presumption of the frequency probability of a phrase, and the frequency probability of a phrase determines the class assignment which follows.
[0047] With this operation gestalt, a forward back WORD algorithm (the FB method) is used as mentioned above for presumption of

the parameter of a Bayh-multi-gram. This is explained in full detail below.
[0048] Using a forward back WORD algorithm, 15 above can make nmax the maximum length of a sequence, and it can calculate T as the number of words of a corpus (text data for study) by the KOMPUREKI city O (nmax2T) which is the degree of complexity. Here, the KOMPUREKI city O (nmax2T) corresponds to the order of calculation cost. That is, the calculation cost concerned of several 15 is proportional to the square of the maximum length nmax of a sequence, and proportional to the number of words of a corpus. In this operation gestalt, fundamentally, it adds not over the set of segmentation {S} but over the time index (t) of a word, and several 15 a molecule and a denominator are calculated. Here, the calculation concerned is dependent on the definition of the variable alpha of the front of the following formula (t, li), and the variable beta of the direction of back (t, lj).
[0049]
[Equation 20] alpha(t, li) =L (W (1) (t-li) | [W (t) (t-li +1)])
[Equation 21] beta(t, lj) =L (W (t+1) (T) | [W (t) (t-lj +1)])
[0050] The variable alpha of front (t, li) expresses the likelihood of the first t words, and the word of the last li individual is restricted here so that one sequence may be formed. moreover, the variable beta of the direction of back (t, lj) -- conditional [of the word of the last individual (T-t)] -- a likelihood is shown and the word of the last individual (T-t) follows a sequence [w(t-lj +1) --w (t)] Here, W (1) and (t-li) express the word train which consists of a word from a time index (1) to (t-li). And if it assumes that the likelihood of an analysis result is calculated by several 7, several 15 will be rewritten like the following formula.
[0051]
[Equation 22] p(k+1) (sj|si) =pc/pd -- here -- [Equation 23]

$$p_c=\sum^{T}\alpha(t,l_i)\,p^{(k)}(s_j\mid s_i)\,\beta(t+l_j,l_j)\,\delta_i(t-l_i+1)\,\delta_j(t+1)$$

t= 1 -- [Equation 24]

$$p_d=\sum_{t}\alpha(t,l_i)\,\beta(t,l_i)\,\delta_i(t-l_i+1)$$

[0052] Here, li and lj show the length of Sequences si and sj, respectively. When the sequence of the word started by the time index t is sk, while Kronecker function deltak (t) is set to 1, when that is not right, it is a function used as 0. Moreover, Variables alpha and beta are calculable with the following repetitive formulas (or recursion formula). Here, in the time index t= 0 and t=T +1, a start and an end symbol are assumed, respectively.
[0053] 1 <=t<=T +1 -- receiving -- : -- [Equation 25]

$$\alpha(t,l_i)=\sum_{l=1}^{n_{max}}\alpha(t-l_i,l)\,p\left([W_{(t-l_i+1)}^{(t)}]\mid[W_{(t-l_i-l+1)}^{(t-l_i)}]\right)$$

It is here and is [Equation 26]. alpha(0 1) = 1, alpha(0 2) = -- It is =alpha(0 nmax) =0.
[0054] 0 <=t<=T -- receiving -- : -- [Equation 27]

$$\beta(t,l_j)=\sum_{l=1}^{n_{max}}p\left([W_{(t+1)}^{(t+l)}]\mid[W_{(t-l_j+1)}^{(t)}]\right)\beta(t+1,l)$$

It is here and is [Equation 28]. beta(1 T+ 1) = 1, beta(1 T+ 2) = -- It is =beta(T+1, nmax) =0.
[0055] When the likelihood of an analysis result is calculated using assumption of a class (i.e., when calculated according to several 19), term [of a re-presumption formula (several-24 22-number)] p (k) and (sj|si) are transposed to (the equivalent of the class, i.e., p(k) (Ck(sj) |Ck(si)) p, (k), and sj|Ck(sj)). the repetitive formula of alpha -- setting -- Term p ([W (t)] (t-li +1) | [W (t-li-l +1) (t-li)]) -- conditional [of the class of a sequence [W (t) (t-li +1)]] -- it is transposed to the BAIGURAMU probability of the corresponding class which carried out the multiplication of the probability Same deformation is performed also about the variable beta in a repetitive formula.
[0056] Subsequently, re-presumption processing in which the forward back WORD algorithm in this operation form was used is explained in full detail below with reference to an example. Instead of a possible analysis result set {S}, addition of several 22 molecule and addition of a denominator perform re-presumption processing of front and the direction of back (henceforth a cross direction) by rearranging two or more terms which can be set to several 15 so that it may be calculated about the time index t of the unit in study data. It depends for this method on the definition of the variable alpha of front, and the variable beta of the direction of back.
(a) Assume that there is no class in following paragraph < <A1>>.
(b) Following paragraph < <A1.1>> defines Variables alpha and beta, and offer an example by it.
(c) Illustrate in following paragraph < <A1.2>> about re-presumption of the cross direction about the frequency probability which used

Variables alpha and beta.
(d) Illustrate about the calculation method of the variables alpha and beta in following paragraph < <A1.3>> repeatedly (or induction).
(e) Following paragraph < <A2>> shows the correction method of paragraph [in case a class exists] < <A1.2>>, and < <A1.3>>.
(f) All the following examples are based on the data shown in the next table.
[0057]
[Table 1]
---------------------------------- Input study data (following) : o n e s

CLAIMS

[Claim(s)]
[Claim 1] Statistical sequence model generation equipment which generates the statistical sequence model of the Bayh-multi-gram which is BAIGURAMU between the natural number N1 piece unit string of variable length, and the natural number N2 piece unit string of variable length based on input data including the sequence which is the unit string which consists of one piece or two or more units which are characterized by providing the following The initialization means which carries out counting of the frequency probability of the above-mentioned motorcycle gram of the combination of all unit strings under the constraint of the maximum of N1 and N2 which were decided beforehand based on the above-mentioned input data The unit string contained in the class classified by classifying into a number of two or more classes which merged so that loss of the mutual information when merging the pair of each class based on the frequency probability of the above-mentioned motorcycle gram by which counting was carried out by the above-mentioned initialization means might serve as the minimum, updated the frequency probability of each class, and were decided beforehand Frequency probability of the conditional unit string of the classified class A classification means to calculate and output the frequency probability of BAIGURAMU between the classified classes, The unit string contained in the classified class which is outputted from the above-mentioned classification processing means, Based on the frequency probability of the conditional unit string of the classified class, and the frequency probability of BAIGURAMU between the classified classes, it is re-presumed using EM algorithm that a maximum likelihood estimate is obtained. here Each unit string of a processing object is received using a forward back WORD algorithm. The front likelihood to the unit string concerned of the processing object serially obtained for the front, The frequency probability of the unit string concerned at the time of being contingent [on the unit string in front of the unit string concerned], By re-presuming the frequency probability of BAIGURAMU between the sequences concerned using the formula showing the frequency probability of BAIGURAMU between sequences based on the back likelihood to the unit string concerned to obtain for back serially A re-presumption means to generate and output the statistical sequence model of the above-mentioned Bayh-multi-gram which it is as a result of re-presumption, and control means controlled to perform processing of the above-mentioned classification means, and processing of the above-mentioned re-presumption means repeatedly until it fulfills predetermined end conditions
[Claim 2] The above-mentioned initialization means is statistical sequence model generation equipment according to claim 1 characterized by removing the data of the combination of BAIGURAMU below predetermined frequency probability further among the frequency probability of BAIGURAMU by which counting was carried out [above-mentioned].
[Claim 3] The above-mentioned classification means is statistical sequence model generation equipment according to claim 1 or 2 characterized by classifying into two or more above-mentioned classes according to the above-mentioned initialization means using the Brown algorithm based on the frequency probability of the above-mentioned motorcycle gram by which counting was carried out.
[Claim 4] The above-mentioned formula the frequency probability of BAIGURAMU between the sequences of a unit string in case the 2nd unit string which is the unit string concerned follows the 1st unit string in the above-mentioned input data It is a formula for calculating to each unit string of the processing object in the above-mentioned input data. the frequency probability of BAIGURAMU between the above-mentioned sequences The claim 1 characterized by being obtained by carrying out the division of the sum of the likelihood in all segmentation containing the 1st and the 2nd unit string by the sum of the likelihood in all segmentation containing the 1st unit string, or statistical sequence model generation equipment of one of 3 publications.
[Claim 5] The denominator which shows the number of times of an average to which each unit string generates the above-mentioned formula in the above-mentioned input data, It has the molecule which shows the number of times of an average to each unit string in case the 2nd unit string follows the 1st unit string in the above-mentioned input data. the above-mentioned molecule It is the sum of the above-mentioned front likelihood to each unit string of a processing object, the frequency probability of the unit string concerned at the time of being contingent [on the unit string in front of the unit string concerned], and the product of the above-mentioned back likelihood. the above-mentioned denominator Statistical sequence model generation equipment according to claim 4 characterized by the above-mentioned front likelihood to each unit string of a processing object, the frequency probability of all the unit strings at the time of being contingent [on the unit string in front of the unit string concerned], and being the sum of the product of the above-mentioned back likelihood.
[Claim 6] The above-mentioned end conditions are the claim 1 characterized by being a time of the number of occurrence of processing of the above-mentioned classification means and processing of the above-mentioned re-presumption means reaching the number of times decided beforehand, or statistical sequence model generation equipment of one of 5 publications.
[Claim 7] It is statistical language model generation equipment which the above-mentioned unit is the character of natural language, the above-mentioned sequence is a word in a claim 1 or the statistical sequence model generation equipment of one of 6 publications, and the above-mentioned classification means classifies a character string into the train of two or more words, and is characterized by the above-mentioned statistical sequence model being a statistical language model.

[Claim 8] It is statistical language model generation equipment which the above-mentioned unit is the word of natural language, the above-mentioned sequence is a phrase in a claim 1 or the statistical sequence model generation equipment of one of 6 publications, and the above-mentioned classification means classifies a word train into the train of two or more phrases, and is characterized by the above-mentioned statistical sequence model being a statistical language model.
[Claim 9] Based on the sound signal of the phonation voice sentence inputted, it is the voice recognition unit characterized by carrying out speech recognition with reference to the statistical language model by which the above-mentioned speech recognition means was generated in the voice recognition unit equipped with the speech recognition means which carries out speech recognition using a predetermined statistical language model with statistical language model generation equipment according to claim 7 or 8.

[Translation done.]